



# Analyzing partially paired data: when can the unpaired portion(s) be safely ignored?

Qianya Qi<sup>a</sup>, Li Yan<sup>b</sup> and Lili Tian<sup>a</sup>

<sup>a</sup>Department of Biostatistics, University at Buffalo, Buffalo, NY, USA; <sup>b</sup>Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA

## ABSTRACT

Partially paired data, either with incompleteness in one or both arms, are common in practice. For testing equality of means of two arms, practitioners often use only the portion of data with complete pairs and perform paired tests. Although such tests (referred as ‘naive paired tests’) are legitimate, their powers might be low as only partial data are utilized. The recently proposed ‘*P*-value pooling methods’, based on combining *P*-values from two tests, use all data, have reasonable type-I error control and good power property. While it is generally believed that ‘*P*-value pooling methods’ are superior to ‘naive paired tests’ in terms of power as the former use more data than the latter, no detailed power comparison has been done. This paper aims to compare powers of ‘naive paired tests’ and ‘*P*-value pooling methods’ analytically and our findings are counterintuitive, i.e. the ‘*P*-value pooling methods’ do not always outperform the naive paired tests in terms of power. Based on these results, we present guidance on how to select the best test for testing equality of means with partially paired data.

## ARTICLE HISTORY

Received 18 November 2019

Accepted 9 December 2020

## KEYWORDS


Hypothesis testing; paired data; normality; *P*-value

## 1. Introduction

Paired data are ubiquitous in medical fields. For example, in genomic experiments of which the purpose is for detecting differentially expressed genes, both cancerous and normal tissues are extracted from each patient. Paired data can eliminate inter-subject variability, hence hypothesis tests using paired samples are generally more powerful than those using unpaired samples. To compare gene expression levels between normal and cancer tissues, a paired test, e.g. paired *t*-test or Wilcoxon signed-rank test, can be performed.

In practice, it is common that not all subjects are able to provide data for both arms, i.e. only a portion of the subjects have both normal and tumor tissues, and the rest have either tumor or normal tissues but not both. The incompleteness in only one arm, say normal arm, yields ‘partially paired data with incompleteness in normal arm’, and the incompleteness in both arms yields ‘partially paired data with incompleteness in both arms.’ In this paper, we assume missing completely at random (MCAR).

**CONTACT** Lili Tian  ltian@buffalo.edu

 Supplemental data for this article can be accessed here. <https://doi.org/10.1080/02664763.2020.1864813>

Let  $X, Y$  denote observations in tumor and normal tissues, respectively. Consider a data set with  $n = n_1 + n_2 + n_3$  subjects where first  $n_1$  subjects provide complete pairs of tumor and normal tissues,  $n_2$  subjects provide only tumor tissues, and  $n_3$  subjects provide only normal tissues. We also assume that  $n_1$  and  $n_2$  are always larger than 0. Therefore, if  $n_3 = 0$ , data is incomplete in normal arm, as shown in Table 1. When  $n_3 > 0$ , we have partially paired data with incompleteness in both arms, as shown in Table 2. Assume that observations of tumor and normal tissues are from populations with means  $\mu_X$  and  $\mu_Y$ , respectively. Let  $\delta = \mu_X - \mu_Y$ . For testing if a gene is up-regulated or down-regulated in tumor samples, we need to test  $H_0 : \delta \leq 0$  against  $H_a : \delta > 0$  or  $H_0 : \delta \geq 0$  against  $H_a : \delta < 0$ . We will focus on the former in this paper.

For testing equality of means in partially paired data, traditionally the most widely used approach is the complete-case analysis, i.e. a paired test using only the paired portion of data with the first  $n_1$  subjects, referred as ‘naive paired test’. This method is straightforward and both parametric and nonparametric paired tests are available in all statistical softwares. Although ‘naive paired test’ is a legitimate method for testing equality of means for partially paired data (i.e. type-I error is controlled), it may have reduced power as it only uses the portion of data with complete pairs. Hence, extensive researches have been conducted targeting using all available data [1–3,5–7,8–19,21,22]. Among them, the combination tests, based on combining  $P$ -values or summary statistics are well studied in literature

**Table 1.** Partially paired data with incompleteness in normal arm.

Subject	Tumor	Normal
1	$X_1$	$Y_1$
2	$X_2$	$Y_2$
3	$X_3$	$Y_3$
$\vdots$	$\vdots$	$\vdots$
$n_1$	$X_{n_1}$	$Y_{n_1}$
$n_1 + 1$	$X_{n_1+1}$	
$n_1 + 2$	$X_{n_1+2}$	
$n_1 + 3$	$X_{n_1+3}$	
$\vdots$	$\vdots$	
$n_1 + n_2$	$X_{n_1+n_2}$	

**Table 2.** Partially paired data with incompleteness in both arms.

Subject	Tumor	Normal
1	$X_1$	$Y_1$
2	$X_2$	$Y_2$
$\vdots$	$\vdots$	$\vdots$
$n_1$	$X_{n_1}$	$Y_{n_1}$
$n_1 + 1$	$X_{n_1+1}$	
$\vdots$	$\vdots$	
$n_1 + n_2$	$X_{n_1+n_2}$	
$n_1 + n_2 + 1$		$Y_{n_1+1}$
$\vdots$		$\vdots$
$n_1 + n_2 + n_3$		$Y_{n_1+n_3}$

due to their reasonable type-I error control and good power properties [1–3,6,7,10,16–18]. Recently, a subclass of combination tests based on combining  $P$ -values, referred as ‘ $P$ -value pooling methods’ hereafter in this paper, were proposed by Kuan and Huang [10] for partially paired data with incompleteness in two arms and Qi *et al.* [16] for partially paired data with incompleteness in single arm. Compared to other methods, ‘ $P$ -value pooling methods’ have overwhelming advantages: (1) great flexibilities in terms of choices of tests, i.e.  $P$ -values can come from any parametric or nonparametric tests; (2) great power property; (3) simple statistical property under null hypothesis; and (4) ease of computation. More details will be given later in this paper or can be found in Kuan and Huang [10] and Qi *et al.* [16].

Despite many statistical methods exist for testing equality of means for partially paired data, ‘naive paired tests’ are still routinely used by practitioners these days due to their simplicity. Therefore, the following questions are intriguing: are ‘naive paired tests’ always inferior to other tests because they only use portion of data?; if not, which settings allow us to use ‘naive paired test’ safely without worrying about losing power? In order to address these questions, we will compare powers by ‘naive paired tests’ and ‘ $P$ -value pooling methods’, i.e. Kuan and Huang [10] for two-arm missing cases, and Qi *et al.* [16] for one-arm missing cases, analytically. We consider settings under normality and with known variance structure (marginal variances and correlation). When the variance structure is unknown, the parameters can be substituted with corresponding consistent estimators and our observations still stand under certain regularity conditions. As paired data with positive correlations are much more common than those with negative correlations in practice, we only focus on scenarios with positive correlations in this paper. More comments and considerations regarding correlation can be found in the last section of this paper.

This paper aims to provide practitioners a general guideline on how to choose between ‘naive paired tests’ and ‘ $P$ -value pooling methods.’ The rest of this paper is organized as follows. Section 2 presents a brief review of ‘ $P$ -value pooling methods’ by Qi *et al.* [16] and Kuan and Huang [10]. In Section 3, power of ‘ $P$ -value pooling methods’ under normality is presented. The results are given in Section 4. Section 5 demonstrates how to use the guidance to choose appropriate methods via some real data examples. Finally, Section 6 gives a summary and discussion.

## 2. Preliminaries

Consider a partially paired data set with incompleteness in either one arm or both arms, as shown in Tables 1 and 2. Let  $(\bar{X}^{(1)}, S_{X^{(1)}}^2)$  and  $(\bar{Y}^{(1)}, S_{Y^{(1)}}^2)$  denote the sample mean and sample variance based on  $n_1$  paired samples for tumor and normal arms, respectively, and let  $S_{X^{(1)}, Y^{(1)}}$  be the sample covariance based on the paired samples. Furthermore, let  $\bar{X}^{(2)}$  and  $S_{X^{(2)}}^2$  be the sample mean and sample variance for the unpaired  $n_2$  tumor samples and  $\bar{Y}^{(2)}$ , and  $S_{Y^{(2)}}^2$  be the sample mean and sample variance for the unpaired  $n_3$  normal samples. We consider testing  $H_0 : \delta = \mu_X - \mu_Y \leq 0$  against  $H_a : \delta > 0$  at significance level  $\alpha$ . Under normality, we aim to compare power of ‘naive paired tests’ with that of ‘ $P$ -value pooling methods’ (i.e. the method by Qi *et al.* [16] for one-arm missing and the method by Kuan and Huang [10] for two-arm missing). In the following, both  $P$ -value pooling methods will be reviewed briefly.

### 2.1. *P-value pooling method for partially paired data with incompleteness in single arm by Qi et al. [17]*

Consider the data structure in Table 1. The gist of  $P$ -value pooling method in [16] is as follows. The paired portion of data  $(X_i, Y_i)$ ,  $i = 1, 2, \dots, n_1$  are used to construct a paired test statistic  $T_p$ , and the unpaired portion of tumor arm  $(X_{n_1+1}, \dots, X_{n_1+n_2})$  and the normal arm of paired portion  $(Y_1, Y_2, \dots, Y_{n_1})$  are used to construct a two-sample test statistic  $T_{up}$ . Let  $P_p$  and  $P_{up}$  be the corresponding  $P$ -values. Also, denote  $F_p$  and  $F_{up}$  as the null distributions for  $T_p$  and  $T_{up}$ . For testing the hypothesis  $H_0 : \delta \leq 0$  vs.  $H_a : \delta > 0$ , the  $P$ -values  $P_p = 1 - F_p(T_p) \sim U(0, 1)$  and  $P_{up} = 1 - F_{up}(T_{up}) \sim U(0, 1)$  under  $H_0$ . Hence, the probit inverse transformations of  $P_p$  and  $P_{up}$ , i.e.  $Z_p = \Phi^{-1}(P_p)$  and  $Z_{up} = \Phi^{-1}(P_{up})$  follow  $N(0, 1)$ . The overall test statistic is defined as:

$$T_1 = \frac{\lambda_1 Z_p + \lambda_2 Z_{up}}{\sqrt{\lambda_1^2 + \lambda_2^2 + 2\lambda_1\lambda_2\eta}}, \quad (1)$$

where  $\eta = \text{corr}(Z_p, Z_{up})$ , and  $\lambda_1$  and  $\lambda_2$  are the weights for the paired test and unpaired test, respectively. Under null hypothesis  $H_0$ ,  $T_1 \sim N(0, 1)$ . Note that the dependence between  $T_p$  and  $T_{up}$ , caused by sharing normal arm  $Y_1, Y_2, \dots, Y_{n_1}$ , results in the dependence between  $Z_p$  and  $Z_{up}$  which will be captured by  $\eta$ .

The power of combination test will be explored using two weighting schemes: (1) unweighted; i.e.  $\lambda_1 = \lambda_2 = 1$ ; (2) weighting by the square root of geometric means of the sample sizes; i.e.  $\lambda_1 = \sqrt{n_1}$ , and  $\lambda_2 = \sqrt{2/(1/n_1 + 1/n_2)}$ . The other weighting schemes explored by Qi et al. [16] include using inverse of standard errors of mean difference estimators and the square root of the sample sizes (i.e.  $\lambda_1 = \sqrt{2n_1}$ ,  $\lambda_2 = \sqrt{n_1 + n_2}$ ). It was discovered that the former could give inflated type I error when sample size is small and the latter yields similar performance as weighting scheme (2) stated above, hence they will not be explored further in this paper.

### 2.2. *P-value pooling method for partially paired data with incompleteness in both arms by Kuan and Huang [11]*

Consider the data structure in Table 2. The gist of  $P$ -value pooling method by Kuan and Huang [10] is given in the following. The paired portion of data  $(X_i, Y_i)$ ,  $i = 1, 2, \dots, n_1$  are used to construct a paired test statistic  $T_p$ , and the unpaired portion of tumor arm  $(X_{n_1+1}, \dots, X_{n_1+n_2})$  and the unpaired portion of normal arm  $(Y_{n_1+1}, Y_{n_1+2}, \dots, Y_{n_1+n_3})$  are used to construct a two-sample test statistic  $T_{up}$ . Obviously  $T_p$  and  $T_{up}$  are independent, so are their  $P$ -values  $P_p$  and  $P_{up}$ . The combination test statistic defined in [10] is

$$T_2 = \frac{\lambda_1 Z_p + \lambda_2 Z_{up}}{\sqrt{\lambda_1^2 + \lambda_2^2}} \sim N(0, 1) \quad \text{under } H_0. \quad (2)$$

Similarly as in 2.1, two weighting schemes (i.e. unweighted and weighting by square root of geometric means of sample sizes) will be explored in power calculation.

### 3. The power

In this section, we compare powers of the ‘*P*-value pooling tests,’ i.e. the method by Qi *et al.* [16] for one-arm missing and the method by Kuan and Huang [10] for two-arm missing, to that of ‘naive paired tests’. A right-sided test is considered, that is,  $H_0 : \delta \leq 0$ , versus  $H_a : \delta > 0$ . Assume  $(X, Y)^T$  follow bivariate normal distributions with mean vector  $\mu = (\mu_X, \mu_Y)^T$  and known covariance matrix  $\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$ .

#### 3.1. Power of the ‘naive paired test’

Despite that the incompleteness is in one arm or both arms, the naive paired test only uses the paired portion of data, i.e.  $(X_i, Y_i)$  where  $i = 1, 2, \dots, n_1$ . With known variances, the paired test statistic is

$$T_p = \frac{\bar{X}^{(1)} - \bar{Y}^{(1)}}{\sqrt{\frac{1}{n_1}(\sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y)}} \sim N(0, 1) \quad \text{under } H_0. \quad (3)$$

The null hypothesis is rejected if  $T_p > z_{1-\alpha}$ , thus the power is

$$\begin{aligned} \text{Power}_p &= \Pr(T_p > z_{1-\alpha} | \delta > 0) \\ &= 1 - \Phi \left( z_{1-\alpha} - \frac{\delta}{\sqrt{\frac{\sigma_Y^2}{n_1}(\nu^2 + 1 - 2\rho\nu)}} \right), \end{aligned}$$

where  $z_{1-\alpha}$  is the  $(1 - \alpha) * 100\%$  quantile of standard normal distribution,  $\Phi(\cdot)$  is the cumulative distribution function of standard normal, and  $\nu^2 = \sigma_X^2/\sigma_Y^2$ .

#### 3.2. Power of *P*-value pooling method with incompleteness in single arm

The paired test  $T_p$  is the same as (3). Using the unpaired portion of tumor arm  $(X_{n_1+1}, \dots, X_{n_1+n_2})$  and the normal arm from the paired portion  $(Y_1, Y_2, \dots, Y_{n_1})$ , given known variances, a two-sample test statistic  $T_{up}$  is defined as

$$T_{up} = \frac{\bar{X}^{(2)} - \bar{Y}^{(1)}}{\sqrt{\frac{\sigma_Y^2}{n_1} \left( \frac{\nu^2}{\gamma_1} + 1 \right)}}, \quad (4)$$

where  $\nu^2 = \sigma_X^2/\sigma_Y^2$ ,  $\gamma_1 = n_2/n_1$ .  $T_{up}$  also follows standard normal distribution under the null hypothesis. Let  $P_p$  and  $P_{up}$  stand for *P*-values from  $T_p$  and  $T_{up}$ , respectively, it is easy to see that  $Z_p = \Phi^{-1}(P_p) = T_p$ ,  $Z_{up} = \Phi^{-1}(P_{up}) = T_{up}$  under  $H_0$ .

The correlation between  $T_p$  and  $T_{up}$  can be easily calculated as

$$\eta = \text{corr}(T_p, T_{up}) = \frac{1 - \rho v}{\sqrt{v^2 + 1 - 2\rho v} \sqrt{v^2/\gamma_1 + 1}}. \quad (5)$$

Substituting  $T_p$ ,  $T_{up}$  and  $\eta$  into (1), we have the combination test statistic  $T_1$  for partially paired data with incompleteness in single arm.

Under null hypothesis,  $T_1$  follows standard normal distribution. Hence the power of  $T_1$  is

$$\begin{aligned} \text{Power}_1 &= \Pr(T_1 > z_{1-\alpha} | \delta > 0) \\ &= \Pr \left( \frac{\lambda_1 T_p - \frac{\lambda_1 \delta}{\sqrt{\frac{\sigma_Y^2}{n_1}(v^2 + 1 - 2\rho v)}} + \lambda_2 T_{up} - \frac{\lambda_2 \delta}{\sqrt{\frac{\sigma_Y^2}{n_1} \left( \frac{v^2}{\gamma_1} + 1 \right)}}}{\sqrt{\lambda_1^2 + \lambda_2^2 + 2\lambda_1 \lambda_2 \eta}} > \right. \\ &\quad \left. z_{1-\alpha} - \frac{\frac{\lambda_1 \delta}{\sqrt{\frac{\sigma_Y^2}{n_1}(v^2 + 1 - 2\rho v)}} + \frac{\lambda_2 \delta}{\sqrt{\frac{\sigma_Y^2}{n_1} \left( \frac{v^2}{\gamma_1} + 1 \right)}}}{\sqrt{\lambda_1^2 + \lambda_2^2 + 2\lambda_1 \lambda_2 \eta}} \right) \\ &= 1 - \Phi \left( z_{1-\alpha} - \frac{\frac{\lambda_1 \delta}{\sqrt{\frac{\sigma_Y^2}{n_1}(v^2 + 1 - 2\rho v)}} + \frac{\lambda_2 \delta}{\sqrt{\frac{\sigma_Y^2}{n_1} \left( \frac{v^2}{\gamma_1} + 1 \right)}}}{\sqrt{\lambda_1^2 + \lambda_2^2 + 2\lambda_1 \lambda_2 \eta}} \right). \quad (6) \end{aligned}$$

To compare powers of  $P$ -value pooling method  $T_1$  and naive paired test  $T_p$ , we define the efficiency function  $f_1(\rho, v, \gamma_1)$  as

$$f_1(\rho, v, \gamma_1) = \frac{\frac{\lambda_1}{\sqrt{(v^2 + 1 - 2\rho v)}} + \frac{\lambda_2}{\sqrt{\left(\frac{v^2}{\gamma_1} + 1\right)}}}{\sqrt{\lambda_1^2 + \lambda_2^2 + 2\lambda_1 \lambda_2 \eta}} * \sqrt{v^2 + 1 - 2\rho v}. \quad (7)$$

If  $f_1(\rho, v, \gamma_1) > 1$ , then  $P$ -value pooling method is more powerful than the naive paired test. The  $f_1(\rho, v, \gamma_1)$  is a complex function of correlation  $\rho$ , the variance ratio of tumor arm to normal arm, i.e.  $v^2 = \sigma_X^2/\sigma_Y^2$ , and the sample size ratio of unpaired tumor sample size  $n_2$  to paired sample size  $n_1$ , i.e.  $\gamma_1 = n_2/n_1$ . Our aim is to find out the values of  $(\rho, v, \gamma_1)$

satisfying  $f_1(\rho, v, \gamma_1) > 1$ . In Section 4, we will present results obtained from numerical calculations.

### 3.3. Power of P-value pooling method with incompleteness in both arms

Similarly to one-arm missing cases, the power of combination test defined in (2) for partially paired data with incompleteness in both arms is

$$\begin{aligned} \text{Power}_2 &= \Pr(T_2 > z_{1-\alpha} | \delta > 0) \\ &= 1 - \Phi \left( \frac{z_{1-\alpha} - \frac{\frac{\lambda_1 \delta}{\sqrt{\frac{\sigma_Y^2}{n_1}(v^2 + 1 - 2\rho v)}} + \frac{\lambda_2 \delta}{\sqrt{\frac{\sigma_Y^2}{n_1} \left( \frac{v^2}{\gamma_1} + \frac{1}{\gamma_1 \gamma_2} \right)}}}{\sqrt{\lambda_1^2 + \lambda_2^2}}}{\sqrt{\lambda_1^2 + \lambda_2^2}} \right), \quad (8) \end{aligned}$$

where  $\gamma_2 = n_3/n_2$ , i.e. the ratio of unpaired normal sample size  $n_3$  to unpaired tumor sample size  $n_2$ , and  $v^2$  and  $\gamma_1$  are the same as defined in 3.2. Define the efficiency function  $f_2(\rho, v, \gamma_1, \gamma_2)$  as

$$f_2(\rho, v, \gamma_1, \gamma_2) = \frac{\frac{\lambda_1}{\sqrt{(v^2 + 1 - 2\rho v)}} + \frac{\lambda_2}{\sqrt{\left( \frac{v^2}{\gamma_1} + \frac{1}{\gamma_1 \gamma_2} \right)}}}{\sqrt{\lambda_1^2 + \lambda_2^2}} * \sqrt{v^2 + 1 - 2\rho v}. \quad (9)$$

The  $f_2(\rho, v, \gamma_1, \gamma_2)$  is a function of  $\rho, v^2, \gamma_1$  and  $\gamma_2$ . When  $f_2(\rho, v, \gamma_1, \gamma_2) > 1$ ,  $P$ -value pooling method is more powerful than the naive paired test. In Section 4, we will present results obtained from numerical calculations.

## 4. Results

In this section, we present the results from extensive numerical studies for comparing powers between the ‘naive paired tests’ and ‘ $P$ -value pooling tests.’ Note that these studies are not simulation studies comparing powers among different tests, instead it is an analytical study investigating how the comparison of powers of these two tests might vary according to parameters. Most importantly, we aim to point out the finding that the ‘naive paired tests’ can be more powerful than ‘ $P$ -value pooling tests’ for certain scenarios. In order to perform a feasible analytical study, data is assumed to follow normal distribution with known variances.

### 4.1. With incompleteness in single arm

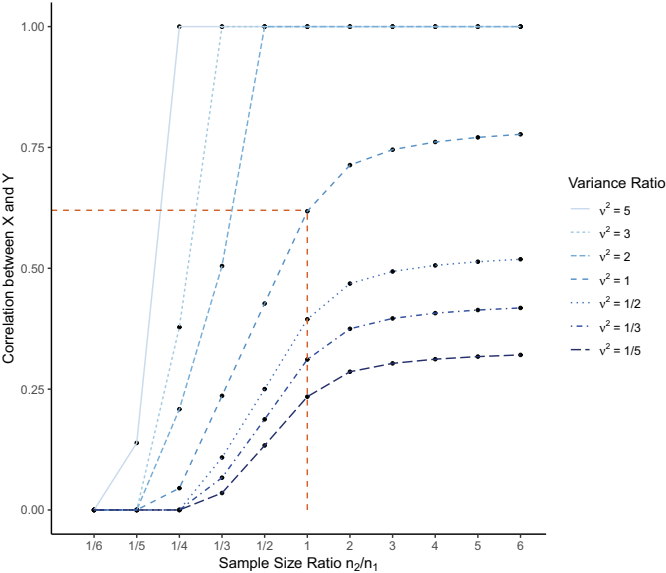
For partially paired data with incompleteness in single arm, we compare the power of the  $P$ -value pooling test by Qi *et al.* [16] and that of naive paired test. As stated in Section 2.1, we consider the  $P$ -value pooling test without and with weights (i.e. square root of geometric

means of the sample sizes). The sample size ratio  $\gamma_1 = n_2/n_1$  is ranging from negatively balanced (1/6, 1/5, 1/4, 1/3, 1/2), balanced (1), to positively balanced (2, 3, 4, 5, 6), and variance ratio  $v^2 = \sigma_X^2/\sigma_Y^2$  is set as 1/5, 1/3, 1/2, 1, 2, 3, 5.

To present an user friendly guideline for choosing the most powerful test to use, Table 3 lists the maximum correlation  $\rho$  for the unweighted and weighted combination tests to be more powerful than the naive paired test given  $\gamma_1$  and  $v^2$ . These values are the same ones used to create Figures 1 and 2.

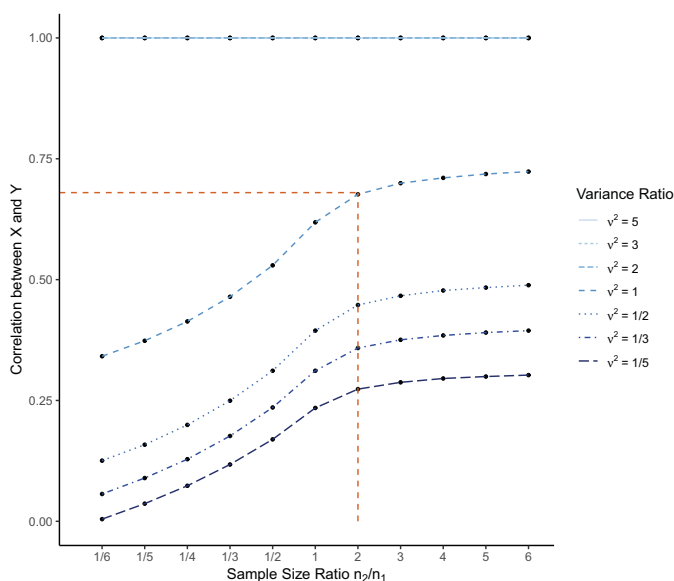
**Table 3.** With incompleteness in single arm: maximum value of correlation  $\rho$  for  $T_1$  (unweighted and weighted) to be more powerful than  $T_p$  given  $\gamma_1 = n_2/n_1$  and  $v^2 = \sigma_X^2/\sigma_Y^2$ .

	$\gamma_1$										
	1/6	1/5	1/4	1/3	1/2	1	2	3	4	5	6
$v^2$	Unweighted										
1/5	0.000	0.000	0.000	0.035	0.134	0.234	0.286	0.303	0.312	0.317	0.321
1/3	0.000	0.000	0.000	0.067	0.188	0.311	0.375	0.396	0.407	0.414	0.418
1/2	0.000	0.000	0.000	0.108	0.250	0.394	0.468	0.493	0.506	0.514	0.519
1	0.000	0.000	0.045	0.236	0.427	0.618	0.714	0.745	0.761	0.771	0.777
2	0.000	0.000	0.208	0.505	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	0.000	0.000	0.378	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5	0.000	0.139	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Weighted										
1/5	0.005	0.037	0.074	0.118	0.170	0.234	0.274	0.288	0.296	0.300	0.302
1/3	0.056	0.090	0.128	0.176	0.236	0.312	0.358	0.376	0.384	0.390	0.394
1/2	0.126	0.158	0.200	0.250	0.312	0.394	0.448	0.466	0.478	0.484	0.488
1	0.342	0.374	0.414	0.464	0.530	0.618	0.677	0.700	0.710	0.718	0.724
2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000



**Figure 1.** With incompleteness in single arm: maximum  $\rho$ 's for unweighted version of combination test  $T_1$  to be more powerful than the naive paired test  $T_p$ , given  $\gamma_1$  and  $v^2$ . The dashed lines correspond to  $\rho = 0.618$  given  $\gamma_1 = 1$ , and  $v^2 = 1$ .





**Figure 2.** With incompleteness in single arm: maximum  $\rho$ 's for weighted version of combination test  $T_1$  to be more powerful than the naive paired test  $T_p$ , given  $\gamma_1$  and  $v^2$ . The dashed lines correspond to  $\rho = 0.677$  given  $\gamma_1 = 2$ , and  $v^2 = 1$ . Note that for  $v^2 = 2, 3, 5$ , the lines overlap.

Figure 1 presents the maximum correlation  $\rho$  (i.e. the maximum correlation for the unweighted  $P$ -value pooling test  $T_1$  being more powerful than the naive paired test  $T_p$ ) vs.  $\gamma_1 = n_2/n_1$ , given different values of  $v^2$ . For example, for homoscedastic and balanced data, i.e.  $v^2 = 1$  and  $\gamma_1 = 1$ , the maximum correlation for the unweighted  $P$ -value pooling test  $T_1$  being more powerful than the naive paired test  $T_p$  is 0.618 (as shown by dashed lines in Figure 1); i.e. the naive paired test is more powerful than the unweighted combination test when the correlation is greater than 0.62. Overall, there are several interesting observations in Figure 1. First of all, given  $v^2 \leq 1$ , the maximum correlation increases as  $\gamma_1$  goes up. For example, given  $v^2 = 1$ , the maximum correlation  $\rho$  is 0.427 when  $\gamma_1 = 1/2$ , and 0.618 when  $\gamma_1 = 1$ . In other words, the naive paired test is more powerful than the combination test when  $\rho > 0.427$  given  $v^2 = 1$  and  $\gamma_1 = 1/2$  and so when  $\rho > 0.618$  given  $v^2 = 1$  and  $\gamma_1 = 1$ . Secondly, when  $\gamma_1 = 1/6$ , i.e. sample size in paired portion is 6-fold of that in unpaired portion, the maximum correlation  $\rho = 0$  despite  $v^2$ , i.e. in terms of power, the combination test is always inferior to the naive paired test. Hence it is always safe to ignore the unpaired portion of tumor arm when its size is less than or equal to one sixth of that of the paired portion despite the other parameters. Thirdly, given  $v^2 = 2, 3, 5$ , the combination test is always superior to the naive paired test (i.e. maximum  $\rho = 1$ ) when sample size is positively balanced.

Figure 2 presents the maximum value of correlation  $\rho$  for the weighted combination test  $T_1$  to be more powerful than the naive paired test  $T_p$  vs. sample size ratio  $\gamma_1 = n_2/n_1$ , given variance ratio  $v^2 = \sigma_X^2/\sigma_Y^2$ . For example, under homoscedasticity, as  $\gamma_1 = 2$ , the maximum correlation for the weighted version of combination test being more powerful than the naive paired test is 0.677; i.e. the naive paired test is more powerful when  $\rho > 0.677$ . From Figure 2, we observe the following: (1) Given  $v^2 \leq 1$ , the maximum correlation increases as

In summary, sample size ratio  $\gamma_1 = n_2/n_1$ , variance ratio  $v^2 = \sigma_X^2/\sigma_Y^2$  and correlation  $\rho$  are the three main factors affecting power and hence our choice of tests.

For partially paired data with incompleteness in both arms, we compare the power of the  $P$ -value pooling test by Kuan and Huang [10] to that of the naive paired test. As stated in Section 2.2, We consider the  $P$ -value pooling test without and with weights (i.e. square root of geometric means of the sample sizes). The first sample size ratio  $\gamma_1 = n_2/n_1$  is ranging from negatively balanced ( $\frac{1}{6}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}$ ), balanced (1), to positively balanced (2, 3, 4, 5, 6); the second sample size ratio  $\gamma_2 = n_3/n_2$  is set as 0.5, 1, 2, 4; and variance ratio  $v^2 = \sigma_X^2/\sigma_Y^2$  is set as  $\frac{1}{5}, \frac{1}{3}, \frac{1}{2}, 1, 2, 3, 5$ .

Figure 3 consists of four subplots which correspond to four settings of  $\gamma_2$ . Each panel presents the maximum values of correlation  $\rho$  for unweighted  $P$ -value pooling test to be more powerful than the naive paired test vs.  $\gamma_1 = n_2/n_1$ , given different values of  $v^2$ . Generally speaking, given  $\gamma_2$  and  $v^2$ , the maximum correlation which allows the unweighted combination test being more powerful than the naive paired test increases as  $\gamma_1$  goes up till  $\gamma_1$  reaches a certain value. However, the relationship between maximum correlation and  $v^2$  is complicated. For example, given  $\gamma_1 = 1/3$  and  $\gamma_2 = 2$ , the maximum

[illegible]

**Table 5.** With incompleteness in both arms ( $\gamma_2 = n_3/n_2 = 1$ ): maximum value of correlation  $\rho$  for combination test  $T_2$  (unweighted and weighted) to be more powerful than the naive paired test given  $\gamma_1 = n_2/n_1$  and  $v^2 = \sigma_X^2/\sigma_Y^2$ .

	$\gamma_1$										
	1/6	1/5	1/4	1/3	1/2	1	2	3	4	5	6
$v^2$	Unweighted										
1/5	0.000	0.190	0.420	0.652	0.882	1.000	1.000	1.000	1.000	1.000	1.000
1/3	0.000	0.164	0.362	0.560	0.758	0.957	1.000	1.000	1.000	1.000	1.000
1/2	0.000	0.150	0.332	0.514	0.697	0.879	0.970	1.000	1.000	1.000	1.000
1	0.000	0.143	0.314	0.486	0.657	0.828	0.915	0.943	0.958	0.966	0.972
2	0.000	0.150	0.332	0.514	0.697	0.879	0.970	1.000	1.000	1.000	1.000
3	0.000	0.164	0.362	0.560	0.758	0.957	1.000	1.000	1.000	1.000	1.000
5	0.000	0.190	0.420	0.652	0.882	1.000	1.000	1.000	1.000	1.000	1.000
	Weighted										
1/5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1/3	0.888	0.892	0.898	0.906	0.922	0.957	1.000	1.000	1.000	1.000	1.000
1/2	0.816	0.820	0.824	0.832	0.846	0.879	0.919	0.943	0.960	0.972	0.980
1	0.768	0.772	0.778	0.784	0.798	0.828	0.866	0.889	0.905	0.916	0.925
2	0.816	0.820	0.824	0.832	0.846	0.879	0.919	0.943	0.960	0.972	0.980
3	0.888	0.892	0.898	0.906	0.922	0.957	1.000	1.000	1.000	1.000	1.000
5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

**Table 6.** With incompleteness in both arms ( $\gamma_2 = n_3/n_2 = 2$ ): maximum value of correlation  $\rho$  for combination test  $T_2$  (unweighted and weighted) to be more powerful than the naive paired test given  $\gamma_1 = n_2/n_1$  and  $v^2 = \sigma_X^2/\sigma_Y^2$ .

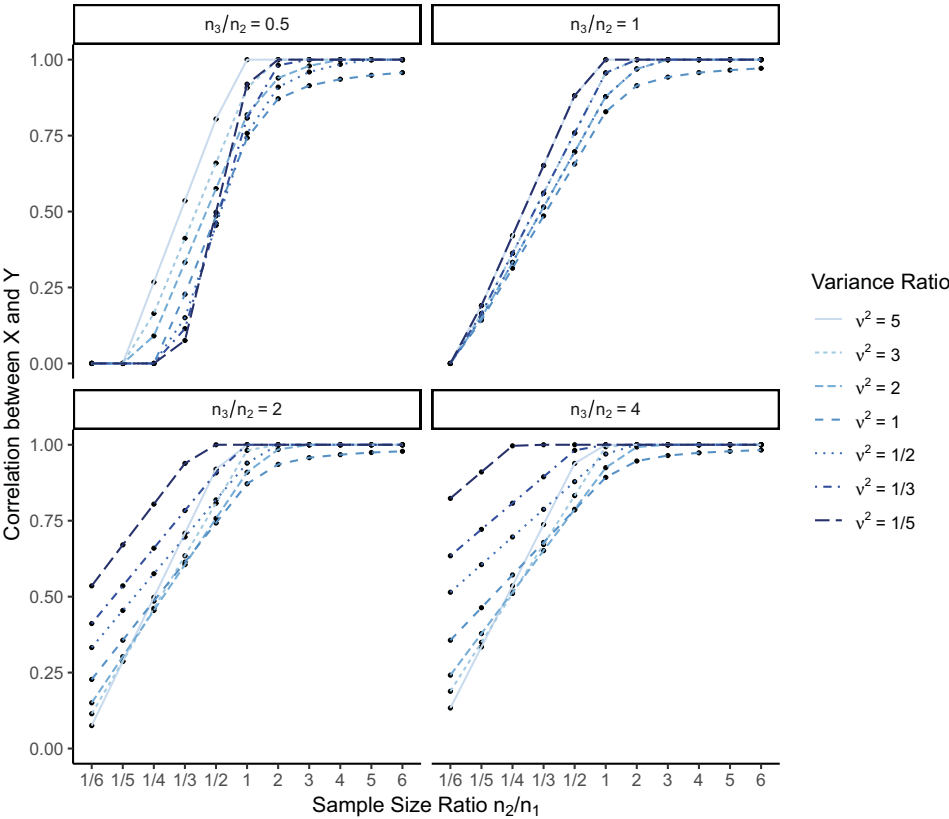
	$\gamma_1$										
	1/6	1/5	1/4	1/3	1/2	1	2	3	4	5	6
$v^2$	Unweighted										
1/5	0.536	0.671	0.804	0.939	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1/3	0.412	0.536	0.660	0.784	0.908	1.000	1.000	1.000	1.000	1.000	1.000
1/2	0.332	0.454	0.576	0.697	0.818	0.940	1.000	1.000	1.000	1.000	1.000
1	0.228	0.356	0.486	0.613	0.742	0.871	0.936	0.958	0.968	0.974	0.978
2	0.150	0.302	0.454	0.605	0.758	0.910	0.984	1.000	1.000	1.000	1.000
3	0.114	0.288	0.462	0.635	0.808	0.982	1.000	1.000	1.000	1.000	1.000
5	0.076	0.287	0.498	0.708	0.920	1.000	1.000	1.000	1.000	1.000	1.000
	Weighted										
1/5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1/3	0.938	0.942	0.948	0.957	0.972	1.000	1.000	1.000	1.000	1.000	1.000
1/2	0.847	0.851	0.857	0.866	0.882	0.914	0.950	0.970	0.984	0.994	1.000
1	0.774	0.778	0.784	0.794	0.810	0.843	0.883	0.905	0.920	0.930	0.938
2	0.794	0.800	0.806	0.818	0.836	0.876	0.922	0.949	0.966	0.978	0.988
3	0.850	0.856	0.864	0.876	0.899	0.944	0.996	1.000	1.000	1.000	1.000
5	0.972	0.978	0.988	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

$\rho = 0.784, 0.613, 0.708$  when  $v^2 = 1/3, 1, 5$ , respectively, indicating a non-monotonic relationship between  $\rho$  and  $v^2$ . Note that when  $\gamma_2 = 1$ , the lines for any given value of  $v^2$  and its reciprocal (e.g.  $v^2 = \frac{1}{2}$  and 2) overlap, as shown in Figure 3.

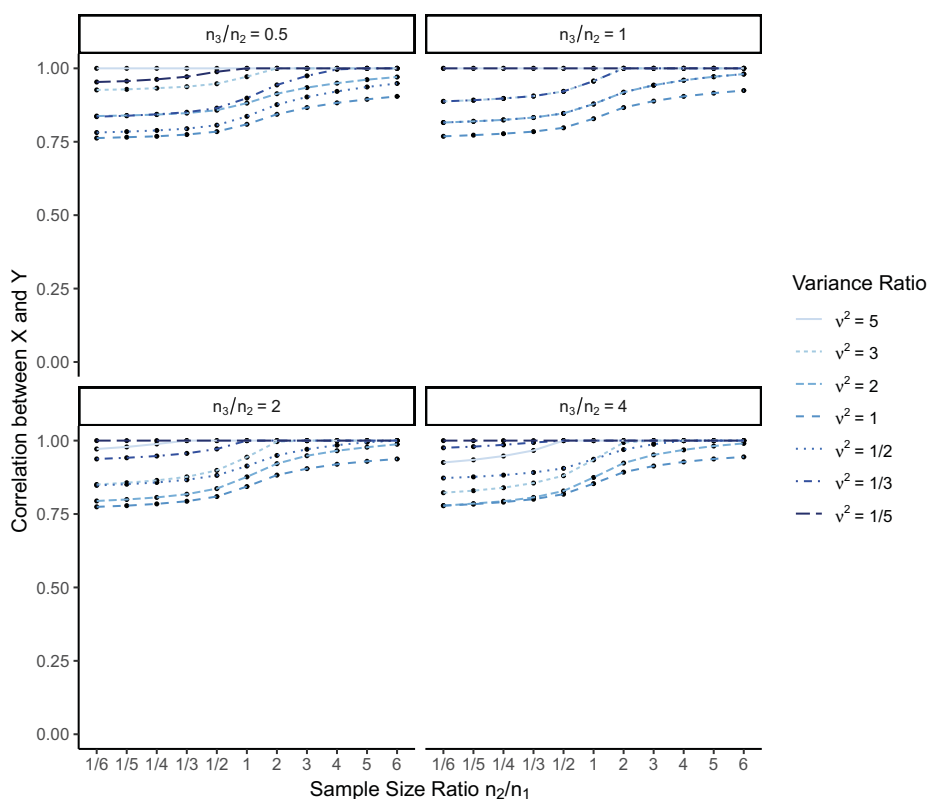
The results for the weighted version of combination test are presented in Figure 4. Comparing to Figure 3, it is clear that the weighted combination test is more powerful than the naive paired test for a bigger range of  $\rho$ . In other words, despite  $\gamma_1, \gamma_2, v^2$ , the maximum  $\rho$  for the weighted combination test being more powerful than the naive paired test is at least 0.75.

**Table 7.** With incompleteness in both arms ( $\gamma_2 = n_3/n_2 = 4$ ): maximum value of correlation  $\rho$  for combination test  $T_2$  (unweighted and weighted) to be more powerful than the naive paired test given  $\gamma_1 = n_2/n_1$  and  $v^2 = \sigma_X^2/\sigma_Y^2$ .

	$\gamma_1$										
	1/6	1/5	1/4	1/3	1/2	1	2	3	4	5	6
$v^2$	Unweighted										
1/5	0.824	0.911	0.996	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1/3	0.635	0.722	0.808	0.895	0.982	1.000	1.000	1.000	1.000	1.000	1.000
1/2	0.514	0.605	0.697	0.788	0.879	0.970	1.000	1.000	1.000	1.000	1.000
1	0.356	0.464	0.572	0.679	0.786	0.893	0.947	0.964	0.974	0.978	0.982
2	0.242	0.378	0.514	0.652	0.788	0.925	0.992	1.000	1.000	1.000	1.000
3	0.188	0.350	0.510	0.672	0.832	0.994	1.000	1.000	1.000	1.000	1.000
5	0.134	0.334	0.536	0.738	0.939	1.000	1.000	1.000	1.000	1.000	1.000
	Weighted										
1/5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1/3	0.976	0.980	0.986	0.994	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1/2	0.872	0.876	0.883	0.892	0.906	0.937	0.970	0.988	1.000	1.000	1.000
1	0.778	0.784	0.790	0.800	0.818	0.853	0.893	0.914	0.928	0.938	0.945
2	0.778	0.786	0.794	0.806	0.828	0.874	0.924	0.952	0.968	0.982	0.990
3	0.822	0.830	0.839	0.855	0.881	0.935	0.994	1.000	1.000	1.000	1.000
5	0.926	0.935	0.948	0.966	1.000	1.000	1.000	1.000	1.000	1.000	1.000



**Figure 3.** With incompleteness in both arms: maximum  $\rho$ 's for unweighted  $T_2$  to be more powerful than the naive paired test  $T_p$  vs.  $\gamma_1$  given  $v^2$ . Each subplot corresponds to a specified value of  $\gamma_2 = n_3/n_2$ .



**Figure 4.** With incompleteness in both arms: maximum  $\rho$ 's for weighted  $T_2$  to be more powerful than the naive paired test  $T_p$  vs.  $\gamma_1$  given  $v^2$ . Each subplot corresponds to a specified value of  $\gamma_2 = n_3/n_2$ .

**Remark:** In Appendix, we consider the asymptotic power of  $P$ -value pooling tests under general settings, i.e. with unknown variances. With incompleteness in one arm, the naive paired test statistic  $T_p$  in (A1) and the two-sample  $t$ -test  $T_{up}$  in (A3) lead to the test statistic  $T_1$  in (A5) of which the asymptotic power, i.e.  $APower_1$  in (A7), converges to  $Power_1$  in (6), under certain regularity conditions. Similar results apply to the scenarios with incompleteness in both arms; i.e. the asymptotic power of  $T_2$ , i.e.  $APower_2$  in (A8), converges to the  $Power_2$  in (8). Additionally, further simulation studies (Tables S1–S4 in ‘Supplemental material’) under normality with unknown variances and under bivariate logistic distribution also demonstrate that the presented results here (Table 3 for incompleteness in one arm and Tables 4–7 for incompleteness in both arms) may be used as crude guidance for choices of tests when dealing with partially paired data.

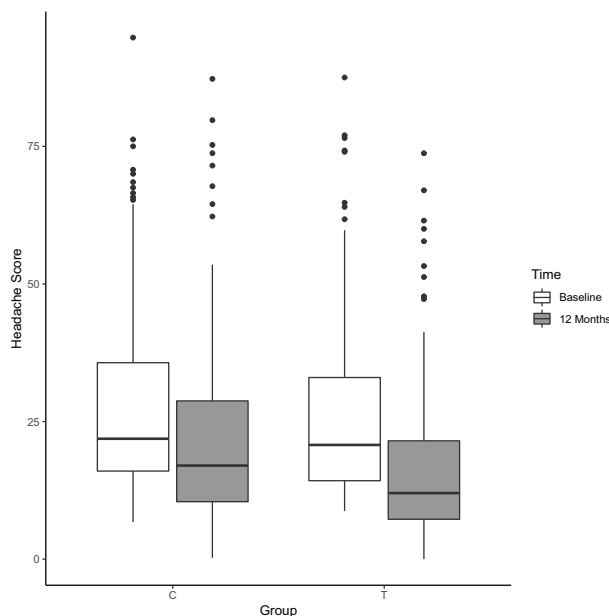
## 5. Real data examples

In this section, we will illustrate how to use Tables 3–7 as guidelines via several real data examples; i.e. how to decide when to ignore the unpaired portion of data and to perform the naive paired test without losing power.

### 5.1. With incompleteness in single arm

**Example 5.1:** Among 90 patients in The Cancer Genome Atlas (TCGA) breast cancer cohort with pathological stage I, 16 of them provided both tumor and normal tissues, and 74 provided only tumor tissues. We are interested in testing whether Gene **ABCC1** is up-regulated, i.e.  $H_0 : \delta \leq 0$  against  $H_a : \delta > 0$ . The sample variances in tumor and normal tissues are 0.344 and 0.162, and the estimated correlation is  $\rho = 0.279$ . Sample size ratio  $\gamma_1 = n_2/n_1 \approx 5$  and variance ratio  $v^2 = \sigma_X^2/\sigma_Y^2 \approx 2$ . According to Table 3, the maximum values of  $\rho$  for  $P$ -value pooling methods to be more powerful than the naive paired test are 1 for both unweighted and weighted tests. Because the estimated correlation  $\rho$  is less than 1, both the unweighted and weighted  $P$ -value pooling methods are more powerful than the naive paired test. Therefore,  $P$ -value pooling tests should be used for testing the hypothesis.

**Example 5.2:** To assess the effects of acupuncture for chronic headache [20], 401 patients were randomly assigned to receive up to 12 acupuncture treatments or to a control intervention offering standard care. The main outcome measures included headache score at baseline and 12 months. Out of the 401 participants, 205 were assigned to treatment group and the rest were assigned to control group, and 44 and 56 participants in the treatment and control groups lost to follow-up at 12 months. We are interested in testing if there is a significant change in headache score from baseline to 12 month for the treatment group and control group separately. Fisher's exact test and logistic regression indicate that the distribution of missing values is not related to either treatment assignment or the headache score at baseline. Thus, MCAR assumption is not violated for this data set. Figure 5 presents



**Figure 5.** Boxplots of headache scores at baseline and 12 months for treatment and control groups for the acupuncture data [20].

the boxplots of headache scores at baseline and 12 months for the treatment and control groups. For the treatment group, the sample size ratio  $\gamma_1 = 0.27$ , the estimated  $\nu^2$  is about 1, and estimated correlation is 0.583. According to Table 3, the maximum correlation for the combination test to be more powerful than the naive paired test is between 0.045 and 0.236 for unweighted test, and between 0.414 and 0.464 for the weighted test. Hence the naive paired test is more powerful than both unweighted and weighted  $P$ -value pooling tests. For the control group, the sample size ratio  $\gamma_1 = 0.40$ , the estimated  $\nu^2$  is about 1, and estimated correlation is 0.811. From Table 3, the maximum correlation for the combination test to be more powerful than the naive paired test is between 0.236 and 0.427 for the unweighted test and between 0.464 and 0.530 for the weighted test. Hence the naive paired test is also more powerful than both unweighted and weighted  $P$ -value pooling tests.

## 5.2. With incompleteness in both arms

**Example 5.3:** To investigate whether the mean Karnofsky score, a patient's functional status measurement, is the same on patients' last two days of life [7], observations of 60 patients were selected to compare the mean difference. Among them, 9 patients provided full data on both days, 28 were scored only on the second to the last day, and 23 were scored only on the last day. The estimated  $\nu^2$  is about 1, the estimated correlation from the paired samples is 0.614, and sample size ratios  $\gamma_1 \approx 3$ ,  $\gamma_2 \approx 1$ . From Table 5, the maximum value of  $\rho$  for the combination test to be more powerful than the naive paired test is 0.943 for unweighted test and 0.889 for the weighted test, respectively. Therefore, both unweighted and weighted  $P$ -value pooling tests are more powerful than the naive paired test for this data set.

**Example 5.4:** To understand the role of the earliest recognizable stages of breast neoplasia in the development of breast cancer, RNAseq libraries were sequenced from formalin-fixed paraffin-embedded tissue of early neoplasia samples and matched normal breast and carcinoma samples from 25 patients [4]. The gene expression levels were compared between normal vs. early neoplasia, normal vs. cancer, and early neoplasia vs. cancer samples. We are interested in testing up-regulation of gene PIK3IP2 between normal and cancer samples. There are 11 patients with complete paired normal and cancer samples, 3 patients with only cancer samples, and 11 with only normal samples. The correlation estimated from the paired samples is 0.833, and sample size ratios  $\gamma_1 \approx 1/4$ ,  $\gamma_2 \approx 4$ , estimated sample variance ratio is about  $\nu^2 = 0.426$ . From Table 7, the maximum value of correlation  $\rho$  for the combination test to be more powerful is between 0.697 and 0.808 for the unweighted test, and between 0.883 and 0.986 for the weighted test. Hence, between unweighted  $P$ -value pooling test and the naive paired test, the latter is a better choice; between weighted  $P$ -value pooling test and the naive paired test, the former is a better choice.

**Remark:** Note that these examples are presented for illustrative purposes and the variances in these examples are estimated but assumed to be known.

## 6. Summary and discussion

Partially paired data is very common in practice. For testing equality of means, practitioners often ignore the unpaired portion(s) and perform 'naive paired tests'. While it is

a common belief that such doing will yield reduced power, a detailed investigation about power comparison has never been done. In this paper, we compare powers of the ‘ $P$ -value pooling tests’ and the ‘naive paired test’ analytically under normality. Our findings are quite counterintuitive, i.e. the ‘naive paired test’ does not suffer from power loss under quite some settings for which the unpaired portion can be safely ignored. Practical guidelines for practitioners are given in Figures 1–4 and Tables 3–7. Furthermore, for data with incompleteness in either single arm or both arms, we observe that the weighted combination test is generally more powerful than the unweighted version. The observation is more obvious when missing data occurs in both arms.

This paper aims to present a counterintuitive point that tests which use all available data do not always outperform the naive paired tests which use only the paired portion of data; i.e. sometimes ‘less is more’. Simulation study is not an efficient way to justify this point as any simulation studies only can cover limited number of scenarios. Therefore, this paper investigates this point analytically in order to make it loud and clear. The power comparison in the paper were performed analytically under normal distribution with known variances. Only with these assumptions, we are able to find the true power under different parameter settings analytically, and are capable of providing simple guidelines such as Tables 3–7.

Although these guidelines are developed under strict assumptions, based on the asymptotic powers and additional simulation studies presented in ‘Supplemental material’, they may serve as crude guidelines in more general cases, i.e. with unknown variances and/or without normality.

This paper only presents settings with positive correlation due to two reasons: (1) positive correlations are much more common for repeated data in practice; (2) when correlation is negative, combination tests are superior to naive paired tests in general in terms of power. This observation agrees with the ‘common sense’, i.e. ‘more data yields higher power’. Therefore, to focus on the aim of this paper, i.e. to present a counterintuitive point that tests using more data does not necessarily yield higher power, analytical results for only positive correlation are presented in the paper.

Out of many existing methods for testing equality of means for partially paired data, this paper focuses on ‘ $P$ -value pooling methods’ by Kuan and Huang [10] for two-arm missing cases, and Qi *et al.* [16] for one-arm missing cases, for power comparison with ‘naive paired tests.’ Besides appealing properties such as good type-I error control and power, the ‘ $P$ -value pooling methods’ also come with great flexibility as the ‘ $P$ -values’ can come from any parametric or nonparametric tests.

Lin and Stivers [11] proposed modified MLE test and several other tests based on simple mean difference estimator for partially paired data with incompleteness in both arms. Via simulation, they touched on the power comparison of their modified MLE test vs. naive paired test. This paper differs from [11] fundamentally: (1) scenarios with incompleteness in one-arm and two-arm are considered; (2) the naive paired test is compared to the ‘ $P$ -value pooling method’ via analytical investigation; (3) the sample size ratio is taken into account in addition to correlation and variance ratio.

Regarding the missing mechanisms, the focus of this paper is MCAR. Our future work will investigate similar problems under MAR (missing at random).

This paper is mainly meant to serve as a reminder to the practitioners and/or researchers that sometimes ‘less is more’ in data analysis.



## Disclosure statement

No potential conflict of interest was reported by the author(s).

## References

- [1] L. Amro, F. Konietzschke, and M. Pauly, *Multiplication-combination tests for incomplete paired data*, Stat. Med. 38 (2019), pp. 3243–3255. arXiv: 1801.08821.
- [2] L. Amro and M. Pauly, *Permuting incomplete paired data: a novel exact and asymptotic correct randomization test*, J. Stat. Comput. Simul. 87 (2017), pp. 1148–1159.
- [3] D.S. Bhoj, *Testing equality of means of correlated variates with missing observations on both responses*, Biometrika 65 (1978), pp. 225–228.
- [4] A.L. Brunner, J. Li, X. Guo, R.T. Sweeney, S. Varma, S.X. Zhu, R. Li, R. Tibshirani, and R.B. West, *A shared transcriptional program in early breast neoplasias despite genetic and clinical distinctions*, Genome Biol. 15 (2014), p. R71.
- [5] G. Ekbohm, *On comparing means in the paired case with incomplete data on both responses*, Biometrika 63 (1976), pp. 299–304.
- [6] Y. Fong, Y. Huang, M.P. Lemos, and M.J. McElrath, *Rank-based two-sample tests for paired data with missing values*, Biostatistics 19 (2017), pp. 281–294.
- [7] B. Guo and Y. Yuan, *A comparative review of methods for comparing means using partially paired data*, Stat. Methods Med. Res. 26 (2017), pp. 1323–1340.
- [8] B.S. Kim, I. Kim, S. Lee, S. Kim, S.Y. Rha, and H.C. Chung, *Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer*, Bioinformatics 21 (2005), pp. 517–528.
- [9] F. Konietzschke, S. Harrar, K. Lange, and E. Brunner, *Ranking procedures for matched pairs with missing data – asymptotic theory and a small sample approximation*, Comput. Stat. Data Anal. 56 (2012), pp. 1090–1102.
- [10] P.F. Kuan and B. Huang, *A simple and robust method for partially matched samples using the p-values pooling approach*, Stat. Med. 32 (2013), pp. 3247–3259.
- [11] P.E. Lin and L.E. Stivers, *On difference of means with incomplete data*, Biometrika 61 (1974), pp. 325–334.
- [12] R.J.A. Little, *Inference about means from incomplete multivariate data*, Biometrika 63 (1976), pp. 593–604.
- [13] S.W. Looney and P.W. Jones, *A method for comparing two normal means using combined samples of correlated and uncorrelated data*, Stat. Med. 22 (2003), pp. 1601–1610.
- [14] P. Martinez-Camblor, N. Corral, and J. Mara de la Hera, *Hypothesis test for paired samples in the presence of missing data*, J. Appl. Stat. 40 (2013), pp. 76–87.
- [15] D.F. Morrison, *A test for equality of means of correlated variates with missing data on one response*, Biometrika 60 (1973), pp. 101–105.
- [16] Q. Qi, L. Yan, and L. Tian, *Testing equality of means in partially paired data with incompleteness in single response*, Stat. Methods Med. Res. 28 (2019), pp. 1508–1522.
- [17] H.M. Samawi and R. Vogel, *Notes on two sample tests for partially correlated (paired) data*, J. Appl. Stat. 41 (2014), pp. 109–117.
- [18] H. Samawi, L. Yu, and R.L. Vogel, *On some nonparametric tests for partially observed correlated data: proposing new tests*, J. Stat. Theory Appl. 14 (2015), pp. 131–155.
- [19] N. Uddin and M.S. Hasan, *Testing equality of two normal means using combined samples of paired and unpaired data*, Commun. Stat. Simul. Comput. 46 (2017), pp. 2430–2446.
- [20] A.J. Vickers, R.W. Rees, C.E. Zollman, R. McCarney, C.M. Smith, N. Ellis, P. Fisher, and R. Van Haselen, *Acupuncture for chronic headache in primary care: large, pragmatic, randomised trial*, BMJ 328 (2004), p. 744.
- [21] J. Xu and S.W. Harrar, *Accurate mean comparisons for paired samples with missing data: an application to a smoking-cessation trial*, Biom. J. 54 (2012), pp. 281–295.
- [22] D. Yu, J. Lim, F. Liang, K. Kim, B.S. Kim, and W. Jang, *Permutation test for incomplete paired data with application to cDNA microarray data*, Comput. Stat. Data Anal. 56 (2012), pp. 510–521.

## Appendix. Asymptotic power under normality with unknown variances

Under normality, when variances are unknown, the naive paired  $t$ -test statistic is

$$T_p = \frac{\bar{X}^{(1)} - \bar{Y}^{(1)}}{\sqrt{\frac{1}{n_1}(S_{X^{(1)}}^2 + S_{Y^{(1)}}^2 - 2S_{X^{(1)}Y^{(1)}})}} \sim t_{n_1-1} \quad \text{under } H_0. \quad (\text{A1})$$

The power is

$$\begin{aligned} \text{Power}_p &= \Pr(T_p > t_{n_1-1;1-\alpha} | \delta > 0) \\ &= 1 - F_T \left( t_{n_1-1;1-\alpha} - \frac{\delta}{\sqrt{\frac{S_{Y^{(1)}}^2}{n_1}(\hat{v}^2 + 1 - 2\hat{\rho}\hat{v})}} \right), \end{aligned}$$

where  $\hat{v}^2$  is the estimated variance ratio  $S_{X^{(1)}}^2/S_{Y^{(1)}}^2$ ,  $\hat{\rho}$  is the estimated correlation  $S_{X^{(1)}Y^{(1)}}/S_{X^{(1)}}S_{Y^{(1)}}$ , and  $F_T$  is the cumulative distribution function of  $t$ -distribution.

As the number of degrees of freedom grows,  $t$ -distribution approaches to standard normal distribution, and since  $\hat{v}$ ,  $\hat{\rho}$  converges in probability to  $v$  and  $\rho$ , respectively, the asymptotic power can be written as

$$A\text{Power}_p = 1 - \Phi \left( z_{1-\alpha} - \frac{\delta}{\sqrt{\frac{\sigma_Y^2}{n_1}(v^2 + 1 - 2\rho v)}} \right), \quad (\text{A2})$$

where  $\Phi(\cdot)$  is the cumulative distribution function of standard normal.

For unpaired data  $(X^{(2)}, Y^{(1)})$ , assuming unequal variances, the two-sample  $t$ -test statistic  $T_{up}$  is defined as

$$T_{up} = \frac{\bar{X}^{(2)} - \bar{Y}^{(1)}}{\sqrt{\frac{S_{Y^{(1)}}^2}{n_1} + \frac{S_{X^{(2)}}^2}{n_2}}}. \quad (\text{A3})$$

Similarly to  $T_p$ , since  $T_{up} \xrightarrow{d} N(0, 1)$ , and  $S_{Y^{(1)}}^2/S_{X^{(2)}}^2$  converges in probability to  $v^2$ ,  $n_2/n_1$  converges to  $\gamma_1$ , the asymptotic power function of  $T_{up}$  is

$$A\text{Power}_{up} = 1 - \Phi \left( z_{1-\alpha} - \frac{\delta}{\sqrt{\frac{\sigma_Y^2}{n_1} \left( \frac{v^2}{\gamma_1} + 1 \right)}} \right). \quad (\text{A4})$$

Let  $P_p$  and  $P_{up}$  stand for P-values from  $T_p$  and  $T_{up}$  respectively, it is easy to see that  $Z_p = \Phi^{-1}(P_p) \rightarrow T_p$ ,  $Z_{up} = \Phi^{-1}(P_{up}) \rightarrow T_{up}$  under  $H_0$ . The test statistic for P-value pooling method in data with incompleteness in single arm is then

$$T_1 = \frac{\lambda_1 Z_p + \lambda_2 Z_{up}}{\sqrt{\lambda_1^2 + \lambda_2^2 + 2\lambda_1\lambda_2\hat{\eta}}}, \quad (\text{A5})$$

where  $\lambda_1$  and  $\lambda_2$  are the weights for the paired test and unpaired test, respectively, and  $\hat{\eta}$  is the asymptotic correlation between  $T_p$  and  $T_{up}$ :

$$\hat{\eta} = \frac{(S_{Y^{(1)}}^2 - S_{X^{(1)},Y^{(1)}})/n_1}{\sqrt{(S_{X^{(1)}}^2 + S_{Y^{(1)}}^2 - 2S_{X^{(1)},Y^{(1)}})/n_1} \sqrt{S_{X^{(2)}}^2/n_2 + S_{Y^{(1)}}^2/n_1}}. \quad (\text{A6})$$

By large sample theory,  $\hat{\eta}$  converges in probability to  $\eta = \frac{1-\rho v}{\sqrt{v^2+1-2\rho v}\sqrt{v^2/\gamma_1+1}}$ . Hence the asymptotic power of  $T_1$  is

$$APower_1 = 1 - \Phi \left( z_{1-\alpha} - \frac{\frac{\lambda_1 \delta}{\sqrt{\frac{\sigma_Y^2}{n_1}(v^2+1-2\rho v)}} + \frac{\lambda_2 \delta}{\sqrt{\frac{\sigma_Y^2}{n_1}\left(\frac{v^2}{\gamma_1}+1\right)}}}{\sqrt{\lambda_1^2 + \lambda_2^2 + 2\lambda_1\lambda_2\eta}} \right). \quad (A7)$$

Thus, the power comparison between P-value pooling method  $T_1$  and naive paired test  $T_p$  is asymptotically equivalent to  $f_1(\rho, v, \gamma_1)$  in manuscript.

For partially paired data with incompleteness in both arms, the two-sample  $t$ -test uses unpaired portion of data  $X^{(2)}$  and  $Y^{(2)}$ , resulting in independent  $T_p$  and  $T_{up}$ . The asymptotic power of  $T_2$  can be written as

$$APower_2 = 1 - \Phi \left( z_{1-\alpha} - \frac{\frac{\lambda_1 \delta}{\sqrt{\frac{\sigma_Y^2}{n_1}(v^2+1-2\rho v)}} + \frac{\lambda_2 \delta}{\sqrt{\frac{\sigma_Y^2}{n_1}\left(\frac{v^2}{\gamma_1} + \frac{1}{\gamma_1\gamma_2}\right)}}}{\sqrt{\lambda_1^2 + \lambda_2^2}} \right), \quad (A8)$$

where  $\gamma_2 = n_3/n_2$ ,  $v^2$  and  $\gamma_1$  are the same as defined in  $APower_1$ . Hence the power comparison between P-value pooling method  $T_2$  and naive paired test  $T_p$  for data with incompleteness in both arms is asymptotically equivalent to  $f_2(\rho, v, \gamma_1, \gamma_2)$  in manuscript.